1 Introduction

Enrichment analysis is a group of approaches for associating a set of genes with a function description. For example, Gene Ontology is one of the most popular gene annotation databases.

Gene Ontology (GO)

Gene ontology (GO) hosts an archive of functional descriptions which are organized in a hierarchical tree structure that depicts the relationships between the terms. Here each term is associated with a parent term that is a little bit more general: cellular component (CC), molecular function (MF) and biological process (BP) are three most general terms. And then the terms are becoming more and more specific as we go down the tree. And then genes can be associated with terms at any level of the tree. And there are many annotators and a consortium that keeps assigning genes to terms.

GO terms are organized hierarchically such that ancestor terms are more general and thus are assigned to more genes, and more specific decendent terms are related to parents by either “is-a” or “part-of” relationships. For example, the nucleus is part of a cell, whereas a neuron is a cell. The relationships form a directed acyclic graph (DAG), where each term can have one or more parents and zero or more children. Users can select the layer of generality the terms capture and carry out analyses accordingly.

Terms are also split into three categories of ontologies:

- Cellular component (CC) - describes where in the cell a gene acts, what organelle a gene product functions in, or what functional complex an enzyme is part of.
• Molecular function (MF) - defines the function carried out by a gene product: one product may carry out many functions; a set of functions together make up a biological process (BP)

• Biological process (BP) - some biological phenomena, or “commonly recognized series of events” affecting the state of an organism. Example include the cell cycle, DNA replication, RNA transcription and translation, limb formation, etc

Besides GO, KEGG is another well-known pathway database for gene annotation.

1.1 Enrichment analysis

Here we will describe some methods for a list of genes or proteins that were identified as differentially expressed or clustered. Or we are concerned only about finding or prioritizing the functional terms based on their enrichment if we have a set of differentially expressed genes or proteins.

So this is the most simple and the most commonly applied test to measure enrichment for a list of genes that were identified experimentally to be prioritize and rank the terms associated with those gene set libraries. So for each set, we compute a p-value that evaluates the enrichment level of that term with your genes of interest.

(1) number of overlapping genes between my list and list from a gene set library
(2) number of genes in the gene set library
(3) number of my differentilly expressed genes
(4) number of all genes

Then, how to compute the enrichment score and the statistic to show how much this observation is deviated from the random situation.

There are currently two major types of approaches for incorporate biological knowledge into differential expression analysis. We will refer to them as the over-representation and aggregate score approaches.

2 Overrepresentation approach

Here we present a contingency table list above, you can fill in the table with your data. And with this table, you can apply Fisher’s exact test to
Table 1: Contingency table

<table>
<thead>
<tr>
<th></th>
<th>$g_i \in S$</th>
<th>$g_i \notin S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(g_i) &lt; .05$</td>
<td>$x$</td>
<td>$n - x$</td>
</tr>
<tr>
<td>$p(g_i) \geq .05$</td>
<td>$D - x$</td>
<td>$(N - D) - (n - x)$</td>
</tr>
</tbody>
</table>

2.1 Fisher’s exact test

Table 2: Contingency table for Fisher’s test

<table>
<thead>
<tr>
<th>a</th>
<th>c</th>
<th>$R_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>d</td>
<td>$R_2$</td>
</tr>
<tr>
<td>$C_1$</td>
<td>$C_2$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

The Fisher’s exact test can compute the $p$-value based upon the above contingency table:

$$p = \frac{\binom{C_1}{a}\binom{C_2}{c}}{\binom{n}{R_1}} = \frac{C_1!C_2!R_1!R_2!}{n!abcld!}$$

2.2 Hypergeometric test

Of total $N$ genes, $n$ pass the threshold, of which $x$ are coming from the gene set $S$ with $D = |S|$:

$$f(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$

The hypergeometric distribution can be approximated by the binomial
with \( p = \frac{D}{N} \):

\[
\lim_{N \to \infty} \binom{D}{x} \binom{N-D}{n-x} \binom{n}{x} = \binom{n}{x} p^x (1-p)^{n-x}
\]

And the cumulative probability can be computed by:

\[
Pr(x > q) = 1 - \sum_{x=0}^{q} \binom{D}{x} \binom{N-D}{n-x} \binom{N}{n}
\]

2.3 Aggregate score approach: GSEA

![GSEA algorithm](image)

Figure 1: GSEA algorithm

Results of the gene set enrichment analysis (GSEA are presented in a format of a table and scatter plots. The analysis was done by using enrichment score (ES) method applied to gene sets with at least 10 genes.

Table 3: GSEA results. In the table, represented are the IDs of the pathways, names of the pathways, number of genes in the pathway, the value of the score statistic, and its significance. Significance infers to the probability of receiving, by chance, a score statistic as extreme as the observed one.

<table>
<thead>
<tr>
<th>Pathway ID</th>
<th>Pathway Name</th>
<th>NGenes</th>
<th>Score</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0003674</td>
<td>molecular function</td>
<td>47</td>
<td>0.538</td>
<td>0.0951</td>
</tr>
<tr>
<td>GO:0050896</td>
<td>response to stimulus</td>
<td>14</td>
<td>0.6307</td>
<td>0.0957</td>
</tr>
<tr>
<td>GO:0050789</td>
<td>regulation of biological process</td>
<td>22</td>
<td>0.5138</td>
<td>0.0959</td>
</tr>
</tbody>
</table>
2.3.1 Inputs to GSEA

(1) Expression data set \( D \in \mathbb{R}^{N \times k} \) with \( N \) genes and \( k \) samples.

(2) Rank the \( N \) genes according to \( a \) priori set ranking metric by some statistical method.

(3) An exponent \( p \) to control the weight of the step.

(4) \( N_H \) Of the \( N \) genes are included in gene set \( S \).

2.3.2 Enrichment score (ES)

(1) Rank the \( N \) genes in \( D \) for form \( L = \{g_1, \ldots, g_N\} \) according to the metric \( r_j = r(g_j) \) of their expression profiles with phenotype \( C \).

(2) Evaluate the fraction of genes in \( S \) ("hits") weighted by their respective metrics, and the fraction of genes not in \( S \) ("misses") present up to a given position \( i \) in \( L \).

\[
P_{\text{hit}}(S, i) = \sum_{g_j \in S, j \leq i} \frac{|r_j|^p}{N_R}
\]

\[
P_{\text{miss}}(S, i) = \sum_{g_j \not\in S, j \leq i} \frac{1}{N - N_H}
\]

where \( N_R = \sum_{g_k \in S} |r_k|^p \) is the summary statistic for gene set \( S \). And therefore the \( ES \) is the maximum deviation from zero of \( P_{\text{hit}} - P_{\text{miss}} \).

For a randomly distributed \( S \), \( ES(S) \) will be relatively small, but if it is concentrated at the top or bottom of the list, or otherwise nonrandomly distributed, then \( ES(S) \) will correspondingly high. When \( p = 0 \), \( ES(S) \) reduces to the standard Kolmorov-Smirnov statistic; when \( p = 1 \), we are weighting the genes in \( S \) by their correlation with \( C \) normalized by the sum of the correlations over all of the genes in \( S \).

2.3.3 Estimating significance

(1) Compute a set of \( ES_{\text{NULL}} \) based on 1000 phenotype permutations.
(2) Compare the observed ES($S$) against the permuted ES$_{NULL}$ values, to obtain the nominal $p$ value for gene set $S$.

### 2.3.4 Multiple hypothesis testing correction

When many gene sets are considered, a correction is performed to account for multiple testing:

1. Determine ES($S$) for each gene set $S$ in the collection.
2. For each set $S$ and each fixed permutation $\pi$ (out of 1000 performed) of the phenotype labels, reorder the genes in $L$ and determine ES($S, \pi$).
3. Adjust for variation in the gene set size:
   
   \[
   \text{NES}(S, \pi) = \frac{\text{ES}(S, \pi)}{\text{MEAN}_{\text{ES}(S, \pi)|\text{ES}(S, \pi)|}}, \text{if } \text{ES}(S, \pi) \geq 0
   \]
   
   \[
   \text{NES}(S) = \frac{\text{ES}(S)}{\text{MEAN}_{\text{ES}(S)|\text{ES}(S)|}}, \text{if } \text{ES}(S) \geq 0
   \]

4. Compute the FDR: create a histogram of all NES($S, \pi$) over all $S$ and $\pi$. Use this null distribution to compute an FDR $q$ value for a given $\text{NES}(S) = \alpha \geq 0$:

   \[q = \frac{|\{(S, \pi)|\text{NES}(S, \pi) \geq \alpha\}|\{(S, \pi)|\text{NES}(S, \pi) \geq 0\}|}{|\{(S, \pi)|\text{NES}(S) \geq \alpha\}|\{(S)|\text{NES}(S) \geq 0\}|}\]

### 3 Conclusion

Enrichment analysis is a means of characterizing biological significance in a given gene set. The GO database provides a central collection of such biological functional annotations assigned to specific genes. The GO ontologies are consisting of cellular component (CC), molecular function (MF), and biological process (BP). With ontologies we can detect functional patterns in a set of genes instead of labeling each gene separately.

GSEA is different from other hypergeometric test based methods, and also offers several advantages. No cutoff is required, and the effects of all genes are taken into account, instead of only a small subset in the latter methods. This, eliminates bias of the choice, but also allows for the possibility of random results showing up as significant. Therefore more corrections need to be made. GSEA also takes into account the strength of each gene’s strength, as apposed to only testing for the membership in specific groups.
4 Some important tools

• Gene Ontology Consortium (GOC): http://geneontology.org/
• DAVID: https://david.ncifcrf.gov/
• GSEA: http://www.broadinstitute.org/gsea/
• Enrichr: http://amp.pharm.mssm.edu/Enrichr